

Muestra representativa de la base del Registro Social de Hogares (RSH)

Contexto

Conforme a lo establecido en el Art. 6 del Decreto Supremo N°22 de 2015, en el marco del Registro Social de Hogares (RSH), la **Subsecretaría de Evaluación Social (SES)** deberá “d) ejecutar los procesos de análisis de la calidad de los datos y del comportamiento de la información contenida en el Registro Social de Hogares”, “g) Administrar el proceso de tratamiento de datos personales de conformidad con lo establecido en el Decreto Supremo N°160 de 2007” y “h) Determinar por medio de uno o más actos administrativos las reglas técnicas de utilización de la información y productos del Sistema”.

En su calidad de administradora del Registro o base de datos, esta Subsecretaría debe dar aplicación a los principios generales contenidos en la Ley N°19.628, sobre protección de la vida privada. En particular, esta minuta describe cómo se busca mantener un equilibrio entre obtener una muestra representativa del país del Registro Social de Hogares (RSH) y proteger la identidad de las personas incluidas en él.

Para lograr una muestra representativa, se implementa un proceso de selección muestral orientado a reproducir adecuadamente la composición de la población inscrita en el Registro Social de Hogares según región y tramo de la calificación socioeconómica (CSE). Paralelamente, se aplican mecanismos de anonimización y protección de datos que resguardan la indeterminación de las personas y hogares incorporados en la muestra, reduciendo el riesgo de reidentificación de los registros publicados.

Metodología de construcción de la muestra

El diseño de la muestra tiene como objetivo garantizar la representatividad y confidencialidad de los datos contenidos en el Registro Social de Hogares (RSH). Para ello, se utiliza una metodología de muestreo aleatorio estratificado, asegurando una adecuada representación de los hogares según región y tramo de calificación socioeconómica (CSE).

Cada estrato corresponde a una combinación única entre región y tramo de calificación socioeconómica, permitiendo representar adecuadamente la diversidad territorial y socioeconómica de la población registrada en el RSH.

La presente muestra fue construida utilizando la información contenida en el Registro Social de Hogares vigente al **1 de junio de 2026**.

Para fines operativos y de resguardo estadístico, se definió una muestra equivalente al 1% del total de hogares contenidos en el Registro Social de Hogares para el período de referencia. Posteriormente, dicho tamaño se distribuye proporcionalmente entre los distintos estratos definidos por región y tramo de calificación socioeconómica.

La unidad de selección corresponde al hogar. En consecuencia, se seleccionan hogares dentro de cada estrato y posteriormente se incorporan todas las personas pertenecientes a los hogares seleccionados.

La selección de hogares se realiza mediante muestreo aleatorio simple sin reemplazo dentro de cada estrato. Sin embargo, en lugar de utilizar una única extracción aleatoria como muestra definitiva, se implementa un procedimiento de simulación múltiple orientado a seleccionar la muestra con mayor similitud respecto de la población total del Registro Social de Hogares.

Para ello, se generan 1.000 muestras independientes, manteniendo constante el diseño estratificado y el tamaño muestral definido para cada estrato. Cada una de estas muestras es evaluada comparando la distribución de un conjunto de variables demográficas y socioeconómicas con las distribuciones observadas para esas mismas variables en la población completa del Registro Social de Hogares.

La comparación entre cada muestra y la población se realiza utilizando el estadístico de Kolmogorov-Smirnov (KS), ampliamente utilizado para medir diferencias entre distribuciones estadísticas. La evaluación considera variables relevantes para la caracterización socioeconómica y demográfica de la población:

- Ingreso
- Ingreso equivalente imputado
- Índice de necesidades
- Edad
- Proporción de hombres

Para cada una de estas variables se calcula de manera independiente el estadístico de Kolmogorov-Smirnov, el cual corresponde a la máxima diferencia observada entre las funciones de distribución acumulada de la población y de la muestra. Valores cercanos a cero indican una mayor similitud entre ambas distribuciones, mientras que valores más altos indican diferencias más pronunciadas. De esta forma, el estadístico KS permite evaluar la capacidad de cada muestra para reproducir no sólo medidas de tendencia central, sino también la forma completa de la distribución observada en la población.

Adicionalmente, se incorpora una medida de diferencia absoluta para la proporción de hombres observada en la muestra respecto de la población total.

Para cada muestra se calcula un indicador agregado de similitud definido como:

$$Score = KS_{y_{def}} + KS_{y_{def_eq_imp}} + KS_{nt_c} + KS_{edad} + D_{sexo}$$

donde:

- $KS_{y_{def}}$ corresponde al estadístico KS para el ingreso.
- $KS_{y_{def_eq_imp}}$ corresponde al estadístico KS para el ingreso equivalente imputado.
- KS_{nt_c} corresponde al estadístico KS para el índice de necesidades del hogar.
- KS_{edad} corresponde al estadístico KS para la edad de las personas.
- D_{sexo} corresponde a la diferencia absoluta entre la proporción de hombres observada en la población y la observada en la muestra.

La muestra definitiva corresponde a aquella que presenta el menor valor del indicador agregado, es decir, aquella cuya distribución conjunta de variables presenta la mayor similitud respecto de la población total del Registro Social de Hogares. Este procedimiento permite reducir la variabilidad inherente a una única extracción aleatoria y seleccionar una muestra que reproduce de mejor forma las características observadas en la población de referencia, manteniendo simultáneamente las propiedades de representatividad propias del diseño estratificado.

Como resultado de este procedimiento, la muestra seleccionada corresponde a **91.782 hogares** y **170.842 personas** pertenecientes a dichos hogares.

Estadísticas descriptivas relevantes: simulaciones múltiples y selección de la muestra

Con el objetivo de evaluar la representatividad de las muestras generadas, se realizaron 1.000 simulaciones independientes manteniendo constante el diseño muestral. Para cada simulación se calcularon estadísticas descriptivas de variables demográficas y socioeconómicas relevantes y posteriormente se compararon con los valores observados en la población total del Registro Social de Hogares.

A partir de estas simulaciones se obtuvo la distribución de los estadísticos muestrales para las variables consideradas en el proceso de selección. Para cada simulación se registraron, entre otros indicadores, la media de las variables continuas y la proporción observada para las variables categóricas de interés.

Con el fin de ilustrar gráficamente los resultados del proceso de selección, se presentan ejemplos correspondientes a dos variables utilizadas en la evaluación de representatividad: edad y proporción de hombres.

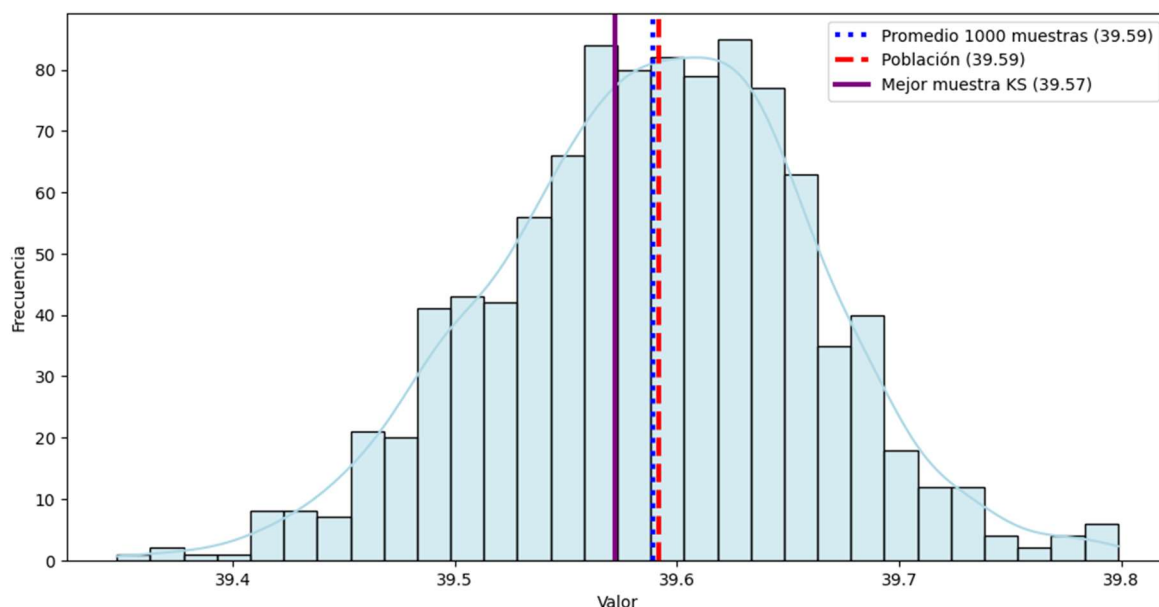
En cada gráfico se representa la distribución de los valores obtenidos en las 1.000 muestras simuladas. Adicionalmente, se incorporan tres referencias de comparación:

- El valor observado en la población total del Registro Social de Hogares.
- El promedio obtenido a partir de las 1.000 muestras simuladas.
- El valor observado en la muestra finalmente seleccionada mediante el criterio de minimización del indicador agregado basado en KS.

La incorporación de estas referencias permite evaluar visualmente el comportamiento del proceso de simulación, verificando que las muestras generadas se distribuyen en torno a los parámetros poblacionales y que la muestra seleccionada reproduce adecuadamente los valores observados en la población de referencia.

El gráfico 1 presenta la distribución de las edades promedio obtenidas en las simulaciones, junto con la edad promedio observada en la población total, el promedio de las 1.000 simulaciones y el valor correspondiente a la muestra finalmente seleccionada.

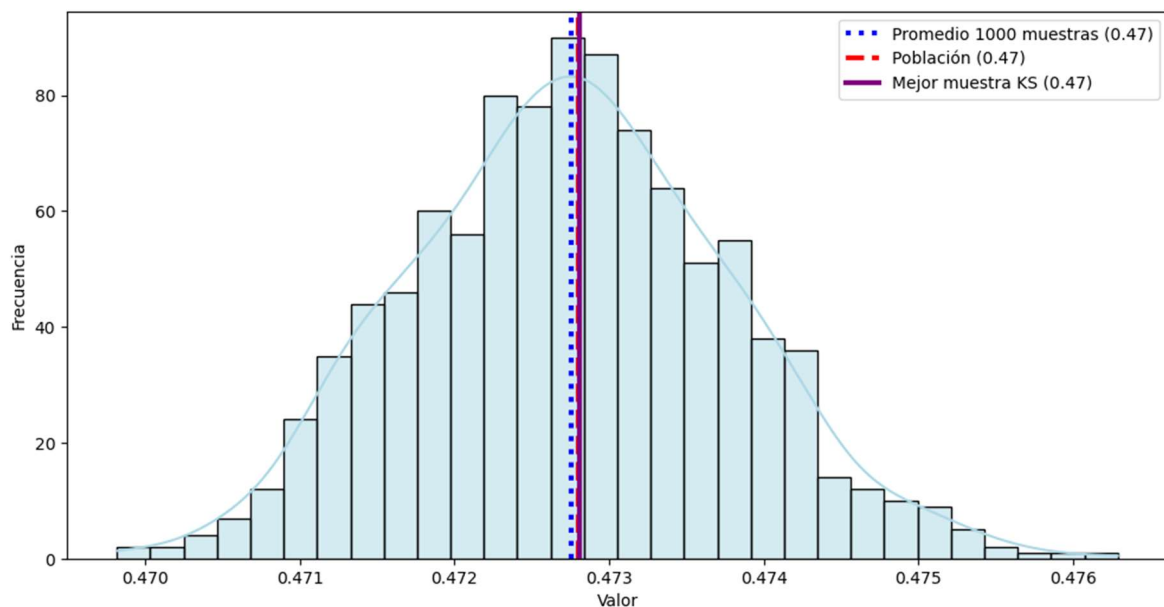
Gráfico 1. Distribución de la edad promedio observada en las 1.000 muestras simuladas.



Fuente: Elaboración Subsecretaría de Evaluación Social en base al Registro Social de Hogares al 1 de junio de 2026.

El gráfico 2 presenta la distribución de la proporción de hombres obtenida en las simulaciones, junto con la proporción observada en la población total, el promedio de las 1.000 simulaciones y el valor correspondiente a la muestra finalmente seleccionada.

Gráfico 2. Distribución de la proporción de hombres observada en las 1.000 muestras simuladas.



Fuente: Elaboración Subsecretaría de Evaluación Social en base al Registro Social de Hogares al 1 de junio de 2026.

Los resultados obtenidos muestran diferencias reducidas entre los valores observados en la población y aquellos presentes en la muestra seleccionada, respaldando la utilización de un procedimiento de simulación múltiple y optimización estadística para definir la muestra definitiva.

Los resultados obtenidos muestran diferencias reducidas entre los valores observados en la población y aquellos presentes en la muestra seleccionada, respaldando la utilización de un procedimiento de simulación múltiple y optimización estadística para definir la muestra definitiva.

Por razones de resguardo de la información y considerando la naturaleza de algunas de las variables utilizadas en el proceso de optimización, los resultados gráficos se presentan únicamente para variables demográficas agregadas. No obstante, el criterio de selección fue aplicado de manera equivalente a todas las variables incorporadas en el cálculo del indicador agregado de similitud.

Actualización de variables y mejoras de calidad de la información

Durante el proceso de actualización de la base muestral se revisaron distintas variables contenidas en la publicación con el objetivo de fortalecer su consistencia respecto de la información disponible en el Registro Social de Hogares.

En particular, se identificó una incidencia en la construcción de la variable asociada a asistencia. Como resultado de la revisión metodológica realizada, se ajustó la construcción de dicha variable para representar de mejor manera la distribución observada para la totalidad de estudiantes presentes en el RSH, independiente de su tipo de dependencia.

Esta mejora incrementa la consistencia de la información publicada respecto de la población total del Registro Social de Hogares y fortalece el potencial analítico de la muestra para estudios vinculados al ámbito educacional.

Limitaciones y recomendaciones de uso

Aunque la metodología implementada busca maximizar la similitud entre la muestra y la población total del Registro Social de Hogares, es importante considerar las siguientes limitaciones y recomendaciones para el adecuado uso de la información.

Representatividad de la muestra

La muestra fue construida utilizando un diseño de muestreo estratificado por región y tramo de calificación socioeconómica, complementado con un proceso de optimización orientado a seleccionar la muestra con mayor similitud respecto de la población total del Registro Social de Hogares. Por esta razón, la muestra presenta una adecuada representación de la estructura poblacional a nivel nacional y para las dimensiones consideradas en el proceso de estratificación y selección. Sin embargo, ello no implica necesariamente representatividad para todos los posibles cruces de variables o para subgrupos de menor tamaño. En particular, los resultados obtenidos para grupos muy específicos de la población deben interpretarse con cautela, especialmente cuando involucren un bajo número de observaciones.

Variables consideradas en el proceso de optimización

La selección de la muestra definitiva se realizó utilizando indicadores asociados a la distribución de ingresos, ingresos equivalentes, índice de necesidades, edad y proporción de hombres. En consecuencia, la similitud entre muestra y población fue optimizada explícitamente respecto de estas variables. Si bien otras variables de la base fueron consideradas durante las etapas de revisión metodológica y validación descriptiva, no forman parte directa del criterio utilizado para seleccionar la muestra definitiva. Por lo tanto, las personas usuarias deben considerar que algunas variables específicas podrían presentar niveles de representatividad distintos de aquellos observados para las variables utilizadas en el proceso de optimización.

Actualización temporal

El Registro Social de Hogares corresponde a una base de datos dinámica que experimenta cambios permanentes producto de nuevas incorporaciones, actualizaciones de información

y modificaciones en la composición de los hogares registrados. En consecuencia, la muestra representa la estructura observada en el Registro Social de Hogares al momento de su extracción y construcción. Cambios posteriores en la población objetivo no se encuentran reflejados en la muestra publicada.

Confidencialidad y protección de datos

Con el fin de resguardar la identidad de las personas y hogares incorporados en la muestra, se implementaron distintos mecanismos de anonimización y agrupación de categorías. Estas medidas buscan reducir el riesgo de identificación directa o indirecta de personas específicas. Como consecuencia, algunas variables presentan niveles de agregación superiores a los disponibles en los registros administrativos originales.

Uso para investigación e inferencia estadística

La muestra fue diseñada principalmente para facilitar el análisis exploratorio, descriptivo y de investigación de la información contenida en el Registro Social de Hogares. La posibilidad de realizar inferencias estadísticas específicas dependerá del objetivo de cada estudio, de las variables analizadas y de los supuestos metodológicos adoptados por cada investigadora o investigador. Por lo tanto, corresponde a cada usuario evaluar la pertinencia de aplicar procedimientos inferenciales sobre la información contenida en la muestra.

Descripción de la estructura de los datos de la muestra

La base de datos del Registro Social de Hogares contiene información proporcionada por un representante de la familia, así como también información obtenida de bases administrativas oficiales de las que dispone el Ministerio. Las variables contenidas caracterizan tanto a las personas como a los hogares de acuerdo con su situación socioeconómica. En la muestra, cada registro corresponde a una persona perteneciente a uno de los hogares seleccionados mediante la metodología descrita previamente.

La estructura de la base de datos considera las siguientes dimensiones:

- **Variables de caracterización de las personas:** identificador ficticio, rango etario, sexo, parentesco en relación con la jefatura del hogar, nacionalidad, pertenencia a pueblo originario, discapacidad, dependencia o necesidades educativas especiales, educación y trabajo.
- **Variables de caracterización del hogar:** identificador ficticio del hogar, tramo de calificación socioeconómica y posesión de medios de alto o muy alto valor.
- **Variables de caracterización de la vivienda:** condiciones de la vivienda, condiciones del sitio, tenencia de la propiedad y disponibilidad de servicios básicos.
- **Variables territoriales:** región y zona de residencia del hogar.

Con el objetivo de proteger la confidencialidad de las personas y la imposibilidad de reidentificación, se aplicaron mecanismos de anonimización y agrupación de categorías en determinadas variables. En general, estas agrupaciones se realizaron sobre categorías de baja frecuencia, permitiendo resguardar la privacidad de las personas sin afectar significativamente la utilidad analítica de la información.

Para comprender mejor el tratamiento realizado a los datos se presenta el siguiente caso. En el formulario original del Registro Social de Hogares la pregunta respecto del parentesco respecto de la jefatura de hogar presenta 13 alternativas de respuestas, que se detallan a continuación:

1. Jefe(a) de Hogar
2. Cónyuge o pareja
3. Hijo(a) de ambos
4. Hijo(a) sólo del jefe(a) de hogar
5. Hijo(a) sólo del cónyuge o pareja
6. Padre o madre
7. Suegro o Suegra
8. Yerno o nuera
9. Nieto(a)
10. Hermano(a)
11. Cuñado(a)
12. Otro familiar
13. No familiar

Sin embargo, se consideró apropiado realizar nuevas agrupaciones. Un caso, reagrupa las opciones de 3 a 5, que corresponden a los hijos que integran una familia. Mientras, que otro caso, las opciones desde la 6 a la 13, que son las alternativas con menor frecuencia queden en una sola categoría.

Las nuevas agrupaciones quedan de la siguiente forma:

1. Jefe de hogar
2. Cónyuge o pareja
3. Hijos/as de ambos o de sólo uno
4. Otros familiares directos o indirecto (Padres / Hermanos / Nietos / Suegros / Yerno- nuera / Cuñados/ Otros)
5. Sin información

De esta forma, se redujo el número de categorías publicadas, manteniendo un adecuado equilibrio entre utilidad estadística y protección de los datos personales.

Para esta publicación, las variables ajustadas corresponden a:

- Rango etario.
- Parentesco.
- Variables de vivienda (tipo de vivienda, situación del sitio, situación de la vivienda, fuente de agua, distribución de agua, sistema de baño, dormitorios ocupados y principal ocupante).

Diccionario de datos de la base muestral

El detalle de los campos, definiciones, tipos de datos, códigos y categorías de respuesta de la base muestral del Registro Social de Hogares se encuentra disponible en el archivo complementario “**Diccionario RSH Muestra**”, publicado en conjunto con este documento metodológico. Dicho archivo forma parte integrante de la documentación técnica de la muestra y constituye la referencia oficial para la interpretación de las variables contenidas en la base de datos.

Se recomienda a las personas usuarias consultar conjuntamente este informe metodológico y el diccionario de variables antes de realizar análisis o procesamiento de la información, con el fin de asegurar una correcta interpretación de los datos publicados.